

Fast computation of Uncertainty Quantification measures in the Geostatistical approach to solve Inverse Problems

Arvind K. Saibaba *

Peter K. Kitanidis *[†]

May 13, 2015

Abstract

We consider the computational challenges associated with uncertainty quantification involved in parameter estimation such as seismic slowness and hydraulic transmissivity fields. The reconstruction of these parameters can be mathematically described as Inverse Problems which we tackle using the Geostatistical approach. The quantification of uncertainty in the Geostatistical approach involves computing the posterior covariance matrix which is prohibitively expensive to fully compute and store. We consider an efficient representation of the posterior covariance matrix at the maximum a posteriori (MAP) point as the sum of the prior covariance matrix and a low-rank update that contains information from the dominant generalized eigenmodes of the data misfit part of the Hessian and the inverse covariance matrix. The rank of the low-rank update is typically independent of the dimension of the unknown parameter. The cost of our method scales as $\mathcal{O}(m \log m)$ where m dimension of unknown parameter vector space. Furthermore, we show how to efficiently compute measures of uncertainty that are based on scalar functions of the posterior covariance matrix. The performance of our algorithms is demonstrated by application to model problems in synthetic travel-time tomography and steady-state hydraulic tomography. We explore the accuracy of the posterior covariance on different experimental parameters and show that the cost of approximating the posterior covariance matrix depends on the problem size and is not sensitive to other experimental parameters.

1 Introduction

One of the central challenges in the field of geosciences is to develop computationally efficient statistical methods for optimizing the use of limited and noisy environmental data to accurately estimate heterogeneous subsurface geological properties. In addition, it is necessary to quantify the corresponding predictive uncertainty. Mathematically, imaging can be performed using inverse problems theory, which uses measurements to make inference of system parameters. Efficient algorithms for inverse problems are necessary to solve problems of realistic sizes, quantified by the spatial resolution of the reconstructed parameters and number of measurements available for reconstruction. Using these efficient algorithms scientists can gain better knowledge of soil moisture content, the porosity of geologic formations, distributions of dissolved pollutants, and the locations of oil deposits or buried liquid contaminants. These detailed images can then be used to better locate natural resources, treat pollution, and monitor underground networks associated with geothermal plants, nuclear waste repositories, and carbon dioxide sequestration sites. We aim to solve these problems by employing the geostatistical approach that stochastically models unknowns as random fields and uses Bayes' rule to infer unknown parameters by conditioning on measurements. However, due to high computational costs in identifying small scale features, these methods are challenging. These

*Department of Electrical and Computer Engineering, Tufts University, 196 Boston Avenue, Medford, MA - 02155
arvind.saibaba@tufts.edu

[†]Department of Civil and Environmental Engineering, Yang & Yamazaki Environment & Energy Building 473
Via Ortega, Stanford, CA 94305 and Institute for Computational and Mathematical Engineering, Huang Engineering
Center 475 Via Ortega, Stanford, CA 94305, peterk@stanford.edu

costs occur because solving inverse problems requires multiple expensive simulations of partial differential equations as well as representing high dimensional random fields, especially on irregular grids and complicated domains. Additional details about the Geostatistical approach have been provided in Section 2.

Uncertainty in the context of Bayesian inverse problems is represented by the posterior probability density function. For linear inverse problems, if the measurement noise is additive Gaussian and the prior model is specified by a Gaussian random field, then the resulting posterior probability density function (PDF) is also Gaussian and is fully specified by calculating the mean (which coincides with the maximum a posteriori, or MAP, estimate), and the posterior covariance matrix. Computing the mean leads to a weighted linear least squares optimization problem, which can be tackled by several efficient numerical algorithms. For nonlinear inverse problems, a linearization of the measurement operator yields a local Gaussian for the posterior PDF. The MAP point can be computed by solving a weighted regularized nonlinear least squares problem and the posterior covariance matrix can be approximated by the inverse of the Hessian of the negative log-likelihood of the posterior PDF computed at the MAP estimate. Although considerable effort has been devoted to computing the MAP estimate (see for example [25, 36, 35]), relatively fewer number of works have addressed the computation of the posterior covariance matrix. In this work we focus on an efficient representation of the posterior covariance matrix which is a measure of uncertainty associated with the reconstruction of the parameters of interest.

Computing and storing the approximation to the posterior covariance matrix is computationally infeasible because the prior covariance matrices arising from finely discretized fields and certain covariance kernels are dense, and computing the dense measurement operator requires solving many forward PDE problems, which can be computationally intractable. In ill-posed inverse problems, the data is informative only about a low-dimensional manifold in the parameter space. This property has been exploited previously in developing efficient approximate representations to the posterior covariance matrix as the sum of the prior covariance matrix and a low-rank update that contains combined information from both the prior and the data misfit part of the Hessian (see for example, [11, 18, 12]). The low-rank modification is computed by the solution of a large-scale eigenvalue problem involving the prior-preconditioned Hessian. The prior covariance matrices can be modeled as discrete representations of operators of the form $\mathcal{A}^{-\alpha}$, where \mathcal{A} is a partial differential operator (for e.g., the Laplacian) and α is a parameter chosen such that the infinite dimensional formulation is well-posed [39]. Another choice for prior covariance matrices is using Spartan Gibbs random field [23]. In this work we focus on the Matérn class of covariance kernels [40].

The ability of being able to compute measures of uncertainty is extremely important for the field of Optimal Experimental Design (OED), which seeks to determine the experimental setups which maximize the amount of information that can be gained about the parameters of interest. The design variables which control the accuracy of the parameter reconstructions could be the measurements or measurement types, numbers, locations of sources and/or detectors and other experimental conditions. A prevalent approach to OED involves optimizing an objective function which involves a scalar measure of uncertainty associated with the parameter reconstruction (defined on the basis of the posterior covariance matrix) and attempts to minimize this objective function with respect to design parameters. Since during the context of optimizing the experimental setup, the inverse problem has to be solved several times and the resulting uncertainty needs to be estimated at each iteration of the optimization routine, we would like an efficient method for computing the objective function (i.e., the measure of uncertainty). We will provide an efficient method for computing a few of these uncertainty measures. A good review of optimal experimental design in the Bayesian context is provided in [13]. Common optimality criteria which can be used as objective functions include the alphabetic criteria, for example, A-, C-, D-, E-, and T- optimality criteria (these will be defined in Section 6). The definitions of the optimality criteria in the geostatistical context, along with a discussion of physical and statistical significance of these criteria and its applicability in non-Gaussian settings is available in [32].

Contributions: We model the prior covariance matrix Γ_{prior} with entries arising from covariance kernels, as is common practice [25]. Although the resulting covariance matrices are dense, in our previous work [36, 4], we have shown that we can obtain the best estimate using techniques (such

as FFT based methods and \mathcal{H} -matrix approach) to reduce the storage and computational cost from $\mathcal{O}(m^2)$ to $\mathcal{O}(m \log m)$ (m is the number of unknown parameters). However, except when the number of measurements are small, e.g., $\mathcal{O}(100)$, computing entries of the posterior covariance matrix is computationally impractical. We show how to compute an efficient representation of the posterior covariance matrix as a low-rank modification of Γ_{prior} and the low-rank update is computed efficiently using a randomized algorithm. A major advantage of our approach is that there is great flexibility in experimenting with several covariance kernels, since the prior covariance matrix computations are handled in a black-box fashion. This is a major difference in our work compared to [11, 18, 12], that we consider directly modeling the entries of the prior covariance matrix using the Matérn class of covariance kernels, instead of modeling the prior as the inverse of a discretized differential operator (such as the Laplacian). Although these two approaches appear disparate, their equivalence has been established in [28].

A second contribution of this paper is that we provide an algorithm for approximating the posterior covariance matrix that does not require forming the square root (or equivalently Cholesky factorization) and inverse of the prior covariance matrix Γ_{prior} . The approach in [11, 18, 12] considers the prior preconditioned Hessian (defined as $\Gamma_{\text{prior}}^{1/2} H_{\text{red}} \Gamma_{\text{prior}}^{1/2}$, where H_{red} is the Gauss-Newton Hessian of the data misfit term). Computing the square root of a matrix is an expensive operation for finely discretized grids arising from large-scale 3D problems. The work in [11, 18, 12] avoids this issue by considering priors for which the square-root is explicitly known. Since the matrix square root is not explicitly known for arbitrary covariance matrices, this assumption is very restrictive from a modeling stand point. The algorithm we propose only requires forming matrix-vector products (henceforth, referred to as matvecs). The key idea is to consider an equivalent generalized eigenvalue problem $Ax = \lambda Bx$ where A is the Hessian corresponding to the data-misfit and $B = \Gamma_{\text{prior}}^{-1}$ is the inverse prior covariance matrix. The randomized algorithm that we propose for approximating the posterior covariance matrix Γ_{post} is simple to implement, is computationally efficient, and comes with error bounds established in our previous work [37].

Another important contribution of our work is the efficient computation of various measures of uncertainty which leverages the efficient representation of the posterior covariance matrix, written as a low-rank correction to the prior covariance matrix. A second computational burden occurs when the number of measurements are large because of operations on a dense cross-covariance matrix, which scale as $\mathcal{O}(n^3)$, where n is the number of measurements. While some of the criteria (A- and C-) can be evaluated when the number of measurements are small, other criteria (such as D- and E-) are altogether computationally infeasible. However, using an efficient representation of the approximate posterior covariance and using matrix-free techniques, we show how several of these optimality criteria can be computed more efficiently. We note a further advantage of using covariance kernels to model Γ_{prior} : Computing the variance of the posterior covariance requires computing the diagonals of the prior covariance matrix which can be easily computed, when the covariance kernel is specified explicitly. The application of computing these uncertainty measures in the context of optimal experimental design has been also covered in [1, 3, 2].

Finally, the efficiency of our proposed algorithms is demonstrated through challenging applications of estimating seismic slowness using traveltime ray tomography (Section 4) and estimating hydraulic conductivity (or transmissivity) using steady state hydraulic tomography (Section 5).

2 Geostatistical approach to inverse problems

The Geostatistical approach (described in the following papers [25, 27, 26]) is one of the prevalent approaches to solve inverse problems. The geostatistical approach represents the unknown field to be estimated as a random field composed as the sum of a few deterministic term, typically low order polynomials, and a stochastic term which models small scale variability. The structure of the stochastic term is represented through the prior probability density function, which in practical applications is often parameterized through variograms and generalized covariance functions. The method has found several applications because it is generally practical and has the ability to quantify uncertainty. The method can generate best estimates, which can be determined in a Bayesian

framework as a posteriori mean values or most probable values, measures of uncertainty as posterior variances or credibility intervals, and conditional realizations, which are sample functions from the ensemble of the posterior probability distribution.

In this section, we briefly review the geostatistical approach. Let $s(\mathbf{x})$, the function to be estimated, be modeled by a Gaussian random field. After discretization, we can write $s \sim \mathcal{N}(X\beta, \Gamma_{\text{prior}})$ where $X \in \mathbb{R}^{m \times p}$ is a matrix of low-order polynomials known as the drift matrix, $\beta \in \mathbb{R}^p$ are a set of drift coefficients to be determined and Γ_{prior} is a covariance matrix with entries $\Gamma_{\text{prior}}(i, j) = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, and $\kappa(\cdot, \cdot)$ is a generalized covariance kernel [15, 30]. We have that $s \in \mathbb{R}^m$ and m refers to the grid size. In several instances, the measurements are corrupted by noise and we model these as Gaussian random variables. The measurement equation can be written as

$$y = h(s) + v \quad v \sim \mathcal{N}(0, \Gamma_{\text{noise}}) \quad (1)$$

where $y \in \mathbb{R}^n$ represents the noisy measurements, $h : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is known as the *measurement operator* or *parameter-to-observation map* and v is a random vector of observation error with mean zero and covariance matrix Γ_{noise} . The matrices Γ_{noise} , Γ_{prior} and X are part of a modeling choice and more details to choose them can be obtained from the following reference [25]. The parameters to be reconstructed are the values of the function at the grid locations s and β which are the drift coefficients.

We can write down the following expressions for the probability density functions (PDF) for the prior and the measurements

$$p(s|\beta) \propto \exp\left(-\frac{1}{2}\|s - X\beta\|_{\Gamma_{\text{prior}}^{-1}}^2\right) \quad (2)$$

and

$$p(y|s) \propto \exp\left(-\frac{1}{2}\|y - h(s)\|_{\Gamma_{\text{noise}}^{-1}}^2\right) \quad (3)$$

where we have used the vector norm $\|x\|_M = \sqrt{x^T M x}$ and M is defined for an arbitrary symmetric positive definite matrix. Inference of the parameters from the measurements is obtained by invoking the Bayes' theorem, through the posterior PDF which is the product of two parts - the likelihood of the measurements and the prior distribution of the parameters. Employing Bayes' rule and assuming that the prior PDF for β is uniform, i.e. $p(\beta) \propto 1$, we can write the expression for the PDF of the posterior distribution as

$$p(s, \beta|y) \propto \exp\left(-\frac{1}{2}\|y - h(s)\|_{\Gamma_{\text{noise}}^{-1}}^2 - \frac{1}{2}\|s - X\beta\|_{\Gamma_{\text{prior}}^{-1}}^2\right) \quad (4)$$

The MAP estimate is computed by maximizing the negative log likelihood of the posterior displayed in Equation (4), and can be obtained by solving the following optimization problem

$$\arg \min_{\hat{s}, \hat{\beta}} \frac{1}{2}\|y - h(s)\|_{\Gamma_{\text{noise}}^{-1}}^2 + \frac{1}{2}\|s - X\beta\|_{\Gamma_{\text{prior}}^{-1}}^2 \quad (5)$$

The choice of prior covariance kernels is the Matérn family of covariance kernels [40]

$$\kappa(\mathbf{x}, \mathbf{y}) = C_{\alpha, \nu}(r) = \frac{1}{2^{\nu-1}\Gamma(\nu)}(\sqrt{2\nu\alpha}r)^\nu K_\nu(\sqrt{2\nu\alpha}r) \quad (6)$$

where $r = \|\mathbf{x} - \mathbf{y}\|_2$, Γ is the Gamma function, α is a scaling factor, $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν . Equation (6) takes special forms for certain parameters ν . For example, when $\nu = 1/2$, $C_{\alpha, \nu}$ corresponds to the exponential covariance function when $\nu = 1/2 + \tilde{n}$ and \tilde{n} is an integer, $C_{\alpha, \nu}$ is the product of an exponential covariance and a polynomial of order \tilde{n} . In the limit as $\nu \rightarrow \infty$, and for appropriate scaling of α , $C_{\alpha, \nu}$ converges to the Gaussian covariance kernel. For a more detailed discussion of permissible covariance kernels, we refer the reader to the following references [30, 15]. Other possible choices for modeling the unknown fields are using the

Spartan Gibbs random field models [23]. Estimating the covariance parameters can be accomplished using the restricted maximum likelihood approach that has been outlined in [25]. The choice of the parameter ν depends on the a priori information available of the smoothness of the field we wish to reconstruct. The effect of the parameters ν has also been studied in Section 4.2. Additional information about the Matérn class can be found in [40].

We discuss the computational costs involving matrix vector products $\Gamma_{\text{prior}}x$ and $\Gamma_{\text{prior}}^{-1}x$. For stationary or translational invariant covariance kernels with points located on a regular equispaced grid, the computational cost for $\Gamma_{\text{prior}}x$ can be reduced from $\mathcal{O}(m^2)$ using the naive approach, to $\mathcal{O}(m \log m)$ by exploiting the connection between Toeplitz structure in 1D or Block-Toeplitz structure in 2D etc, and the Fast Fourier Transform (FFT) [31]. For irregular grids, it can be shown that the cost for approximate matrix-vector products (matvecs) involving the prior covariance matrix Γ_{prior} can be reduced to $\mathcal{O}(m \log m)$ using Hierarchical matrices [36] or $\mathcal{O}(m)$ using \mathcal{H}^2 -matrices or Fast Multipole Method (FMM) [4]. For forming matvecs $\Gamma_{\text{prior}}^{-1}x$, we use an iterative solver such as GMRES with a preconditioner that employs approximate cardinal functions based on local centers and special points [7]. The cost of constructing the preconditioner is $\mathcal{O}(m)$ or $\mathcal{O}(m \log m)$. Assuming that the number of iterations is independent of the size of the system, the cost of forming $\Gamma_{\text{prior}}^{-1}x$ is also $\mathcal{O}(n_{\text{iter}}m)$ or $\mathcal{O}(n_{\text{iter}}m \log m)$, where n_{iter} is the number of iterations required to converge to the desired tolerance. In conclusion, the cost for forming $\Gamma_{\text{prior}}x$ and $\Gamma_{\text{prior}}^{-1}x$ is $\mathcal{O}(m \log m)$.

In the case that the operator $h(s) = Hs$ is linear, the resulting posterior PDF in Equation (4) is also Gaussian. The MAP point (5) can be calculated by introducing auxillary variables $\hat{\xi}$ and $\hat{\beta}$ which satisfy the following system of equations

$$\begin{pmatrix} H\Gamma_{\text{prior}}H^T + \Gamma_{\text{noise}} & HX \\ (HX)^T & 0 \end{pmatrix} \begin{pmatrix} \hat{\xi} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix} \quad (7)$$

The MAP estimate can be calculated as $\hat{s} = X\hat{\beta} + \Gamma_{\text{prior}}H^T\hat{\xi}$. Further details on the $\hat{\xi}$ - $\hat{\beta}$ formulation and the procedure to compute the solution to the linear system of equations in Equation (7) can be found in [36, 35]. Furthermore, the posterior covariance matrix is given by the expression

$$\Gamma_{\text{post}} \stackrel{\text{def}}{=} \begin{pmatrix} F_{ss} & F_{s\beta} \\ F_{s\beta}^T & F_{\beta\beta} \end{pmatrix} = \begin{pmatrix} \Gamma_{\text{prior}}^{-1} + H^T\Gamma_{\text{noise}}^{-1}H & \Gamma_{\text{prior}}^{-1}X \\ X^T\Gamma_{\text{prior}}^{-1} & X^T\Gamma_{\text{prior}}^{-1}X \end{pmatrix}^{-1} \quad (8)$$

3 Posterior Covariance approximation

In [18, 11], the approach taken to approximate the posterior covariance is to compute the dominant modes of a prior preconditioned data misfit Hessian $\Gamma_{\text{prior}}^{1/2}H_{\text{red}}\Gamma_{\text{prior}}^{1/2}$, which combines the information from the data misfit portion of the Hessian H and the prior covariance matrix Γ_{prior} . Here, $H_{\text{red}} \stackrel{\text{def}}{=} -\nabla_{ss} \log p(y|s) = H^T\Gamma_{\text{noise}}^{-1}H$, where $p(y|s)$ is the likelihood of the measurements conditioned on the data. The matrix H_{red} often has a rapidly decaying spectrum for several ill-posed inverse problems [18, 11, 9, 10]. Computing the dominant modes of this eigenvalue problem can be accomplished in a cost that is a small multiple of the cost of simulating the forward (or adjoint) problem, and this multiple is independent of the size of the parameter dimension. Once this low-rank approximation is computed, an approximation to the posterior covariance matrix can be computed as a low-rank update to the prior covariance matrix using the Woodbury matrix identity. This results in a scalable algorithm for estimating uncertainty in large-scale statistical inverse problems.

3.1 Avoiding $\Gamma_{\text{prior}}^{1/2}$ and $\Gamma_{\text{prior}}^{-1}$

As mentioned earlier, the key drawback of [18, 11] is that the algorithms for approximating posterior covariance matrix involves computing (matvecs with) the square root of the prior covariance matrix. When the prior covariance matrix is represented by a discrete representation of a (possibly fractional) differential operator [39], the square root can be computed using sparse Cholesky representation that scales as $\mathcal{O}(m^{3/2})$ in 2D and $\mathcal{O}(m^2)$ for 3D problems.

Several matrix-free techniques exist in the literature for computing matvecs $\Gamma_{\text{prior}}^{1/2}x$ and $\Gamma_{\text{prior}}^{-1/2}x$, that are based on polynomial approximation [14, 17] or rational approximations and contour integrals [20]. However, the convergence of polynomial approximations is only algebraic when the smallest eigenvalue is close to zero. Rational approximations and contour integral based methods do not suffer from the same problem, however they require solutions of a number of shifted systems totaling $\mathcal{O}(\log \kappa_{\text{cond}}(\Gamma_{\text{prior}}))$, where $\kappa_{\text{cond}}(\cdot)$ is the condition number defined for a matrix A as $\kappa_{\text{cond}}(A) \stackrel{\text{def}}{=} \|A\|_2 \|A^{-1}\|_2$. Although the number of systems to be solved is often small, even for ill-conditioned problems, solving each system can be expensive in practice. Recent developments also include symmetric square-root factorizations, such as \mathcal{H} -Cholesky [8]. However, they are complicated to implement and have low accuracy for a reasonable computational cost. Therefore, it is highly desirable to develop square-root free algorithms for approximating the posterior covariance matrix.

Our approach, instead, is to consider the dominant eigenmodes of the generalized Hermitian eigenvalue problem (henceforth referred to as GHEP) $H_{\text{red}}x = \lambda \Gamma_{\text{prior}}^{-1}x$, which has the same spectrum as the matrix prior preconditioned data misfit Hessian $(\Gamma_{\text{prior}}^{1/2}H_{\text{red}}\Gamma_{\text{prior}}^{1/2})$. However, as we shall show in Section 3.2 the dominant eigenmodes can be computed without relying on square-roots of the matrix Γ_{prior} . We consider the generalized eigenvalue problem

$$H^T \Gamma_{\text{noise}}^{-1} H u = \lambda \Gamma_{\text{prior}}^{-1} u \quad (9)$$

Recall that $H_{\text{red}} = H^T \Gamma_{\text{noise}}^{-1} H$. Since both matrices H_{red} and Γ_{prior} are symmetric and Γ_{prior} is symmetric positive definite, we have the following eigendecomposition

$$H_{\text{red}} = \Gamma_{\text{prior}}^{-1} U \Lambda U^T \Gamma_{\text{prior}}^{-1} \quad U^T \Gamma_{\text{prior}}^{-1} U = I \quad (10)$$

where U is the matrix of eigenvectors obtained by the solution of Equation (9) and Λ is a diagonal matrix with eigenvalues. Equation (9) does not have square-roots but still has $\Gamma_{\text{prior}}^{-1}$. We further make the following variable transformation, $y = \Gamma_{\text{prior}}^{-1}x$. Therefore, Equation (9) becomes

$$\Gamma_{\text{prior}} H^T \Gamma_{\text{noise}}^{-1} H \Gamma_{\text{prior}} y = \lambda \Gamma_{\text{prior}} y \quad (11)$$

Note that this transformation still preserves the same eigenvalues λ . After solving for the eigenpair (λ, y) the eigenvector of Equation (9) can be recovered by the following transformation $x = \Gamma_{\text{prior}} y$. As a result, we have completely removed the need for inverting Γ_{prior} or forming square roots. We now consider efficient solvers for computing the dominant eigenmodes of the GHEP (9).

3.2 A randomized algorithm for approximating posterior covariance

In Section 3.1, we described an efficient representation of the posterior covariance matrix, based on the dominant eigenmodes of the eigenvalue problem described in Equation (9). Equation (9) is a GHEP, it is a special case of a generalized eigenvalue problem $Ax = \lambda Bx$ with A being Hermitian and B being Hermitian positive definite. Several efficient methods exist for solving the GHEP. These include approaches based on power and inverse iteration methods, Lanczos based methods and Jacobi-Davidson method. For a good review on this material, please refer to [6, chapter 5] and [34]. A good review of existing software available for solving eigenvalue problems is also available in [22].

In this paper, we choose to use the square-root free randomized algorithms that we have recently developed in [37]. Randomized algorithms are gaining popularity because 1) they are easy to implement, 2) they are well suited for computationally intensive problems and modern computing environments, and 3) have well-established error bounds and computational costs. The error analysis suggests that the randomized algorithm developed in [37] is most accurate when the generalized singular values of the matrix $B^{-1}A$ decay rapidly. In Section 3.1, we have argued that the generalized eigenvalues of the GHEP (9) are rapidly decaying and therefore, we expect the randomized algorithm to be fairly accurate. The randomized algorithms that we employ are also extremely efficient because they avoid forming expensive matvecs with $B^{1/2}$ and $B^{-1/2}$. In addition, they produce a symmetric low-rank representation with B-orthonormal eigenvectors. This symmetry in

low-rank representation has been exploited in deriving an efficient, symmetric representation of the posterior covariance matrix in Section 3.3.

We briefly review the randomized algorithm described in [37] for computing dominant eigenmodes of the GHEP $Ax = \lambda Bx$. In the context of solving the problem (9), we have

$$A \stackrel{\text{def}}{=} \Gamma_{\text{prior}} H^T \Gamma_{\text{noise}}^{-1} H \Gamma_{\text{prior}} \quad \text{and} \quad B \stackrel{\text{def}}{=} \Gamma_{\text{prior}}$$

The key observation is that the matrix $C \stackrel{\text{def}}{=} B^{-1}A = H_{\text{red}}\Gamma_{\text{prior}}$ is symmetric with respect to the B -inner product $\langle x, y \rangle_B = y^T Bx$. Suppose we wanted to compute the k largest generalized eigenpairs of $Ax = \lambda Bx$. The randomized Algorithm 1 calculates a matrix Q , which is B -orthonormal and approximately spans the column space of C , i.e. satisfies the following error bound $\|(I - QQ^T B)C\|_B \leq \varepsilon$. Given such a matrix Q , it can be shown that $\|A - (BQ)(Q^T A Q)(BQ)^T\|_B \leq 2\varepsilon$, i.e. $A \approx (BQ)(Q^T A Q)(BQ)^T$. As a result, a symmetric rank- k approximation can be computed, from which the approximate eigendecomposition can be computed.

To produce a symmetric rank- k approximation, the algorithm proceeds as follows: first, we sample a matrix with entries randomly chosen from $\mathcal{N}(0, 1)$, $\Omega \in \mathbb{R}^{n \times r}$. We choose $r = k + p$, where p is an oversampling factor, which we choose to be 20. Numerical evidence for this choice of oversampling factor has been provided in [21, 37]. Form $\bar{Y} = A\Omega$ and $Y = B^{-1}A\Omega$. Then, we compute the QR factorization of $Y = QR$ such that $Q^T BQ = I$. This can be accomplished by modified Gram-Schmidt algorithm with $\langle \cdot, \cdot \rangle_B$ inner products. Then, we form $T \stackrel{\text{def}}{=} Q^T A Q$ and compute its eigenvalue decomposition $T = S\Lambda S^T$. We then have the approximate generalized eigendecomposition

$$A \approx U\Lambda U^T \quad U = QS$$

Here, U is also B -orthonormal. The cost of a second round of matvecs with A while computing T can be avoided using the following approximation

$$\Omega^T \bar{Y} = \Omega^T A \Omega \approx (\Omega^T BQ)T(Q^T B\Omega)$$

Therefore, T can be computed as $T \approx (\Omega^T BQ)^{-1}(\Omega^T \bar{Y})(Q^T B\Omega)^{-1}$. The results are compactly summarized in Algorithm 1. For an efficient implementation we describe the algorithm using $A = \Gamma_{\text{prior}} H^T \Gamma_{\text{noise}}^{-1} H \Gamma_{\text{prior}}$ and $B = \Gamma_{\text{prior}}$ and $B^{-1}A = H^T \Gamma_{\text{noise}}^{-1} H \Gamma_{\text{prior}}$.

Algorithm 1 Randomized algorithm for computing dominant eigenmodes of $H_{\text{red}}x = \lambda \Gamma_{\text{prior}}^{-1}x$

Require: matrices $H_{\text{red}}, \Gamma_{\text{prior}} \in \mathbb{R}^{n \times n}$ and $\Omega \in \mathbb{R}^{n \times (k+p)}$ is a Gaussian random matrix. Here, k is the desired rank, $p \sim 20$ is an oversampling factor.

- 1: Compute $Y = H_{\text{red}}\Gamma_{\text{prior}}\Omega$
- 2: Compute QR - factorization of $Y = QR$ s.t. $Q^T \Gamma_{\text{prior}} Q = I$
- 3: Form $T = Q^T \Gamma_{\text{prior}} H_{\text{red}} \Gamma_{\text{prior}} Q$ or $(\Omega^T BQ)^{-1}(\Omega^T \bar{Y})(Q^T B\Omega)^{-1}$
- 4: Compute the eigenvalue decomposition $T = S\Lambda S^T$. Keep the k largest eigenmodes.
- 5: **Return:** Matrices $U \in \mathbb{R}^{n \times k}$ and $\Lambda \in \mathbb{R}^{k \times k}$ that satisfy

$$H_{\text{red}} \approx (\Gamma_{\text{prior}}^{-1} U) \Lambda (\Gamma_{\text{prior}}^{-1} U)^* \quad \text{with} \quad U = \Gamma_{\text{prior}} Q S$$

The efficiency and accuracy of this algorithm has been studied in detail in [37]. Here we summarize the main conclusions. The error in the low-rank approximation is

$$\|(I - QQ^T B)B^{-1}A\|_B \leq c\|B^{-1}\|_2 \sigma_{B,k+1}(B^{-1}A)$$

where c is a constant that depends on n, k and p and is independent of the spectrum of the matrices, $\sigma_{B,k+1}$ is the $(k+1)$ -th generalized singular value of the matrix $B^{-1}A$. Since the generalized singular values are not known in advance, a randomized estimator for the error in the low-rank representation is also proposed and analysed in [37]. Given error in the low-rank representation

$\|(I - QQ^TB)B^{-1}A\|_B$, it can be shown that the error in approximating the true eigenvalue and eigenvector satisfies the following error bounds

$$|\lambda - \tilde{\lambda}| \leq \min\{2\varepsilon, \frac{4\varepsilon^2}{\delta}\} \quad \sin \angle_B(u, \tilde{u}) \leq \frac{2\varepsilon}{\delta}$$

where $\delta = \min_{\lambda_i \neq \lambda} |\tilde{\lambda} - \lambda_i|$ is the gap between the approximate eigenvalue $\tilde{\lambda}$ and any other eigenvalue and $\angle_B(x, y) = \arccos \frac{|\langle x, y \rangle_B|}{\|x\|_B \|y\|_B}$. This result states that the accuracy in the eigenvalue/eigenvector calculations depends not only on the accuracy of the low-rank representations but also on the spectral gap δ . When the eigenvalues are clustered, the spectral gap is small and the eigenvalue calculations are accurate as long as the error in the low-rank representation is small. However, in this case the resulting eigenvector calculations may be inaccurate because the parameter δ appears in the denominator for the approximation of the angle between the true and approximate eigenvector. This result has consequences in the accuracy approximation of the posterior covariance and will be discussed in Section 4.2.

3.3 Posterior covariance approximation

Let us define the matrix $F_{ss} \stackrel{\text{def}}{=} \left(H^T \Gamma_{\text{noise}}^{-1} H + \Gamma_{\text{prior}}^{-1} \right)^{-1}$. We follow the approach in [18, 29] to derive an approximation to F_{ss}^{-1} as a low-rank update to the prior covariance matrix Γ_{prior} . Plugging in the approximate eigendecomposition for H_{red} from Equation (10), we obtain

$$F_{ss} = \left(\Gamma_{\text{prior}}^{-1} U \Lambda U^T \Gamma_{\text{prior}}^{-1} U + \Gamma_{\text{prior}}^{-1} \right)^{-1}$$

Using the Sherman-Morrison-Woodbury formula, we can derive the following expression

$$\begin{aligned} F_{ss}^{-1} &= \Gamma_{\text{prior}} - \Gamma_{\text{prior}} \Gamma_{\text{prior}}^{-1} U \left(\Lambda^{-1} + U^T \Gamma_{\text{prior}}^{-1} U \right)^{-1} U^T \Gamma_{\text{prior}}^{-1} \Gamma_{\text{prior}} \\ &= \Gamma_{\text{prior}} - U (\Lambda^{-1} + I)^{-1} U^T \\ &= \Gamma_{\text{prior}} - U_k D_k U_k^T + \mathcal{O} \left(\frac{\lambda_{k+1}}{1 + \lambda_{k+1}} \right) \end{aligned} \quad (12)$$

where $D_k \stackrel{\text{def}}{=} \text{diag}(\frac{\lambda_i}{1 + \lambda_i}) \in \mathbb{R}^{k \times k}$ for $i = 1, \dots, k$. The error in the approximation to the matrix F_{ss} and its inverse can be established by the following inequalities

$$\begin{aligned} \|F_{ss} - \left(\Gamma_{\text{prior}}^{-1} + \Gamma_{\text{prior}}^{-1} U_k \Lambda_k U_k^T \Gamma_{\text{prior}}^{-1} \right)\|_M &\leq \lambda_{k+1} \|\Gamma_{\text{prior}}^{-1}\|_2 \\ \|F_{ss}^{-1} - (\Gamma_{\text{prior}} - U_k D_k U_k^T)\|_M &\leq \frac{\lambda_{k+1}}{1 + \lambda_{k+1}} \|\Gamma_{\text{prior}}\|_2 \end{aligned} \quad (13)$$

The inequality holds for both the cases that $M = \Gamma_{\text{prior}}$ and $M = \Gamma_{\text{prior}}^{-1}$. Here, the induced matrix norm is defined as $\|A\|_M \stackrel{\text{def}}{=} \max_{\|x\|_M=1} \|Ax\|_M$, defined for symmetric positive definite matrices M .

For many ill-posed inverse problems, the approximate numerical rank k of the prior preconditioned Hessian $\Gamma_{\text{prior}}^{1/2} H_{\text{red}} \Gamma_{\text{prior}}^{1/2}$ is small and independent of the problem size, i.e. the number of unknowns. Since this matrix has the same eigenvalues as the GHEP $H_{\text{red}} x = \lambda \Gamma_{\text{prior}}^{-1} x$, we are justified in retaining only the eigenmodes corresponding to the largest eigenvalues. We can exploit this low-rank structure to develop scalable algorithms to approximate the posterior covariance matrix. The generalized eigendecomposition combines information from the prior and the reduced Hessian and takes advantage of the eigenvalue decay in one (or both) matrices - when the reduced Hessian has rapidly decaying eigenvalues, or the prior is of smoothing type. Analytical evidence for the eigenvalue decay of the reduced Hessian H_{red} is provided in [18] in the context of advection-diffusion based inverse problems and in [9, 10] for inverse scattering problems. For the case of prior covariance matrices, the eigenvalue spectrum is known to decay rapidly when the covariance kernels are

smooth [38]. The k retained eigenvectors are the modes along which the parameter field is informed by a combination of the data and the prior. Typically the data and prior are informative about the low-frequency modes and as a result local information and fine scale information is hard to recover from the data. The inequalities in (13) suggests choosing k such that $\lambda_{k+1} \ll 1$.

Finally, the posterior covariance matrix can be constructed by using the following decomposition in Equation 8

$$\begin{aligned}\Gamma_{\text{post}} &= \begin{pmatrix} I & -F_{ss}^{-1}F_{s\beta} \\ & I \end{pmatrix} \begin{pmatrix} F_{ss}^{-1} & \\ & (F_{\beta\beta}^{-1} - F_{s\beta}^T F_{ss}^{-1} F_{s\beta})^{-1} \end{pmatrix} \begin{pmatrix} I & \\ -F_{s\beta}^T F_{ss}^{-1} & I \end{pmatrix} \\ &= \begin{pmatrix} F_{ss}^{-1} + GS^{-1}G^T & -GS^{-1} \\ -S^{-1}G^T & S^{-1} \end{pmatrix}\end{aligned}\quad (14)$$

where $S \stackrel{\text{def}}{=} (F_{\beta\beta}^{-1} - F_{s\beta}^T F_{ss}^{-1} F_{s\beta})$ and $G \stackrel{\text{def}}{=} F_{ss}^{-1} F_{s\beta}$.

An approximation to the posterior covariance matrix (14) can be constructed by plugging in the approximation to F_{ss}^{-1} which has been derived in Equation (12). It should be noted that computing and storing the posterior covariance matrix or its inverse is infeasible for very large problem sizes. However, neither computing nor storing the matrix Γ_{post} is necessary, nor recommended. Computing $F_{ss}^{-1}x$ can be performed as

$$F_{ss}^{-1}x \approx \Gamma_{\text{prior}}x - U_k D_k U_k^T x \quad \text{Cost : } \mathcal{O}(m \log m + km)$$

By the same argument, forming G and S is approximately the same cost as forming matrix-vector product with F_{ss}^{-1} . Having computed G and S , the matrix-vector product involving Γ_{post} is dominated by the cost of computing $F_{ss}^{-1}x$ and can be computed efficiently in $\mathcal{O}(m \log^\gamma m + rm)$.

4 Application: Ray Tomography

4.1 Setup

In this application, we consider a synthetic setup of cross-well tomography [4]. The goal is to the image the slowness in the medium where slowness is defined as the reciprocal of seismic velocity. As a first order approximation, the seismic wave is modeled as traveling along a straight line from the sources to the receivers without reflections or refractions. Each source-receiver pair generates one measurement and therefore, there are $n = n_{\text{rec}} n_{\text{sou}}$ measurements. Here n_{rec} is the number of receivers and n_{sou} are the number of sources. In this application, we pick $n_{\text{sou}} = 20$ and $n_{\text{rec}} = 50$. The domain is discretized into m cells $\sqrt{m} \times \sqrt{m}$ and within each cell, the slowness is assumed to be constant. Therefore, the time taken from the source to the receiver is a weighted sum of the slowness in the cell, weighted by the length of the ray within the cell. The inverse problem can be stated as follows: given measurements of time delay between each source-receiver pair, what is the slowness of the medium? We assume that the travel times are corrupted by Gaussian noise, so that the measurement takes the following form

$$y = Hs + v \quad v \sim \mathcal{N}(0, \Gamma_{\text{noise}}) \quad (15)$$

where y are the observed (synthetic) travel times, s is the slowness that we are interested in imaging and H is the measurement operator, whose rows correspond to each source-receiver pair and are constructed such that their inner product with the slowness would result in the travel time. Constructed as above results in H as a sparse matrix with $\mathcal{O}(n\sqrt{m})$ non-zero entries - each row has $\mathcal{O}(\sqrt{m})$ entries and there are n measurements. As a ‘‘true field’’, we take a realization from a Gaussian random field $s_{\text{true}} \sim \mathcal{N}(X\beta_{\text{true}}, \Gamma_{\text{prior}})$, with $X = [1, \dots, 1]^T$, $\beta_{\text{true}} = 5 \times 10^{-3}[s/m]$ and covariance matrix Γ_{prior} is constructed according to the covariance kernel

$$\kappa(\mathbf{x}, \mathbf{y}) = \theta \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2}{L}\right) \quad (16)$$

with $\theta = 1 \times 10^{-3}[\text{s/m}]$ and $L = 100[\text{m}]$. The realization of the stationary Gaussian process is generated using the FFT-based approach in [16] by embedding the Toeplitz matrix in a circulant matrix. The matvecs with Γ_{prior} are also performed with FFT based methods in $\mathcal{O}(m \log m)$.

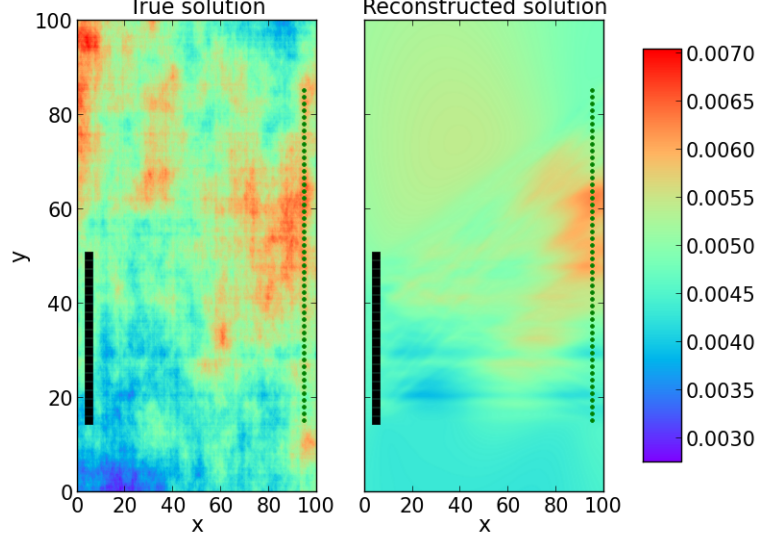


Figure 1: (left) True field which is a realization of a Gaussian process with mean $5 \times 10^{-3}[\text{s/m}]$ and covariance kernel (16), and (right) Reconstruction using the geostatistical approach on a grid with 1024×1024 points. Black solids correspond to source locations (20) and green dots correspond to the receiver locations (50). The reconstruction error in relative L^2 sense was 0.11.

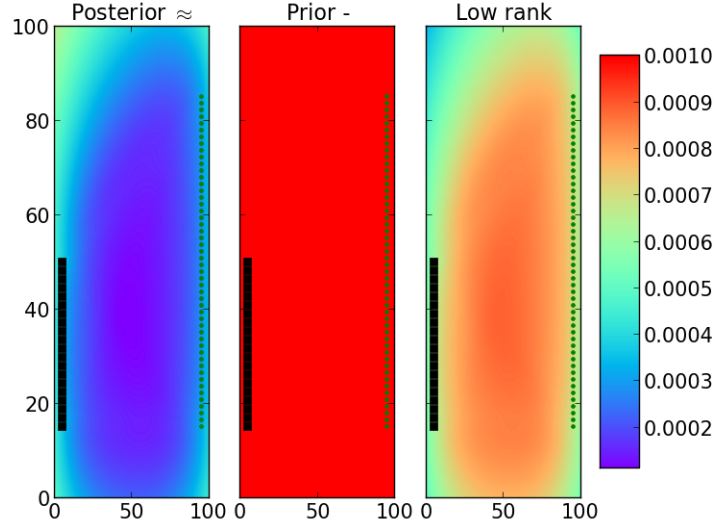


Figure 2: Visualization of the diagonals of the posterior covariance matrix. (left) diagonals of the posterior covariance matrix, (middle) diagonals of the prior covariance matrix and (right) the low-rank correction from the (1,1) block in Equation (14). The uncertainty in the estimation of β measured as $\text{Trace}(S^{-1})$ (see Equation (8)) is $6.32 \times 10^{-4} [\text{s/m}]$.

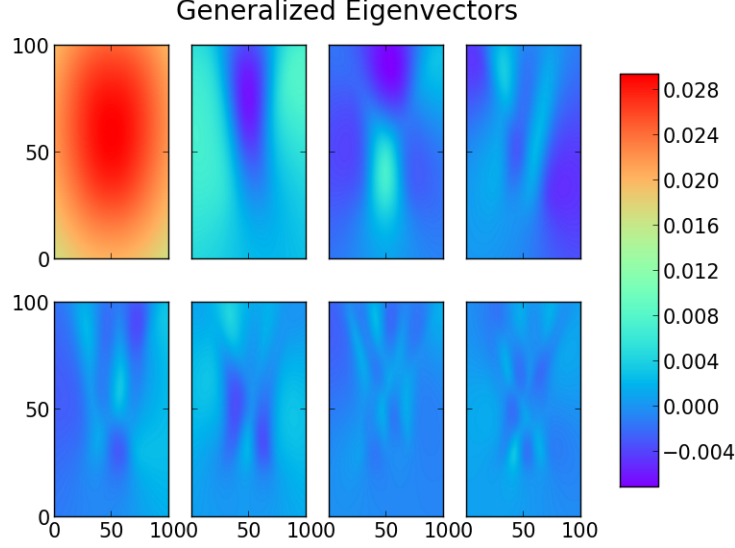


Figure 3: Eigenvectors 1, 3, 5, 7, 10, 12, 18 and 20 of the generalized eigenvalue problem (10) solved using Algorithm 1.

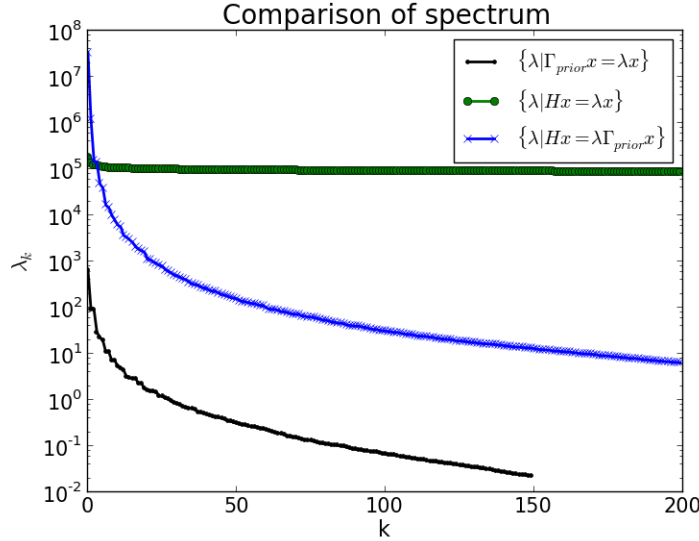


Figure 4: Comparison of the eigenvalues of the data misfit Hessian $H_{\text{red}} = H^T \Gamma_{\text{noise}}^{-1} H$, the prior covariance matrix Γ_{prior} and the GHEP in the Equation (9). As can be seen clearly, the eigenvalues of the GHEP that combines information from both the prior Γ_{prior} and the data misfit Hessian H_{red} decay rapidly. The grid size used in this calculation was 1024×1024 .

To simulate experimental error, we add Gaussian noise of 0.1% to the measurements. The system of Equations (7) is solved using an iterative solver, which is chosen to be restarted GMRES (50), and the matvecs can be formed in a matrix-free manner [36, 35]. As a preconditioner, we use the method developed in [36, Section 3.3]. The preconditioner requires computing a low-rank decomposition of Γ_{prior} which can be computed using any eigenvalue solver. We use the randomized GHEP algorithm described in 3.2 with $A = \Gamma_{\text{prior}}$ and $B = I$, with ~ 150 eigenpairs used to form the low-rank decomposition and an oversampling factor $p = 20$. The other steps in the

construction of the preconditioner are identical. To simulate experimental error, we add Gaussian noise of 0.1% to the measurements. The iterative solver converged in 138 iterations. The L^2 error in the reconstruction was 11.08%. Figure 1 shows the true field which is a realization of a Gaussian process with mean 5×10^{-3} [s/m] and covariance kernel (16) and the reconstructed field obtained by solving Equation (7). The grid size was 1024×1024 and the relative reconstruction error in the l_2 sense was 0.11. Figure 2 visualizes the posterior variance obtained as the diagonals of the posterior covariance matrix calculated using the formula $\Gamma_{\text{post}} = \Gamma_{\text{prior}} - U_k D_k U_k^T$. The error in the uncertainty associated with estimation of $\hat{\beta}$ measured as $\text{Trace}(S^{-1})$ (see Equation (8)) is 6.32×10^{-4} [s/m]. The visualization a few selected eigenvectors of the generalized eigenvalue problem (10) solved using Algorithm 1 can be seen in Figure 3.

4.2 Variance approximation

In this section, we study the effect of truncation of the spectrum and the choice of prior on the accuracy of the variance of the reconstruction, that is the diagonals of the posterior covariance matrix. We consider three different covariance kernels chosen from the Matérn class of covariance kernels corresponding to parameters $\nu = 1/2, 3/2, 5/2$, i.e.

$$\kappa_\nu(\mathbf{x}, \mathbf{y}) = \begin{cases} \theta \exp(-r/L) & \nu = 1/2 \\ \theta(1 + \sqrt{3}r/L) \exp(-\sqrt{3}r/L) & \nu = 3/2 \\ \theta(1 + \sqrt{5}r/L + 5(r/L)^2) \exp(-\sqrt{5}r/L) & \nu = 5/2 \end{cases} \quad (17)$$

The comparison of the rate of decay of the eigenvalues of the matrices Γ_{prior} , Γ_{post} and the GHEP $H_{\text{red}}x = \lambda \Gamma_{\text{prior}}^{-1}x$ is illustrated in Figure 4. What does affect the spectrum of the generalized eigenvalue problem for a given measurement setup is the choice of the covariance kernel used in the reconstruction. Here we illustrate this by choosing a ray tomography setup with $n_m = 400$ measurements and unknowns discretized on a 200×200 grid. For the Matérn class of covariance kernels, by increasing the parameter ν , the underlying stochastic process becomes more smooth and the rate of decay of the eigenvalues increases [40, 38]. This affects the number of eigenvalues retained to maintain the accuracy of the posterior covariance matrix - smoother the kernel, fewer eigenvalues need to be retained. The number of retained eigenvalues is $\min\{n_m, k(\varepsilon)\}$, where $k(\varepsilon)$ is chosen such that the retained eigenvalues satisfy the cutoff $\lambda > \varepsilon$. A typical choice is $\varepsilon = 0.1$. This is illustrated in Figure 5. We see that for the covariance kernel corresponding to $\nu = 1/2$, the eigenvalue decay is not rapid enough to satisfy the cutoff and as a result, all the non-zero eigenvalues that correspond to the number of measurements $n_m = 400$. Physical parameters also control the decay of eigenvalues and as a result, the number of eigenmodes that are retained. Setups that have more number of measurements due to increased number of sources and receivers, have a slower rate of decay of eigenvalues because more information is propagated through the rows of the measurement operator. This is also illustrated in Figure 5.

4.2.1 Accuracy of variance approximation

In Figure 6, we plot the relative error in the computation of the approximate variance as a function of the number of eigenvalues retained for different prior covariance kernels. The error is defined to be the relative error computed with the exact variance. We observe that the error decreases with increasing number of eigenvalues retained. Consequently, inclusion of a larger number of eigenvalues presents a situation of diminishing returns in terms of the accuracy of variance calculations. The result of the computation of the accuracy of the variance with increasing eigenvalues is plotted in Figure 6. As can be seen, the error in the approximation to the variance decreases rapidly with increasing number of eigenvalues retained. Furthermore, the accuracy improves dramatically when considering kernels with higher values of ν , since eigenvalues of covariance matrices arising from the Matérn family corresponding to larger ν decay more rapidly. In our computations, we assumed that the number of measurements are only 400, this makes the task of computing the exact variance computationally feasible using the method described in [4]. For larger number of measurements, we cannot compute the variance easily and we must resort to the methods described in this paper.

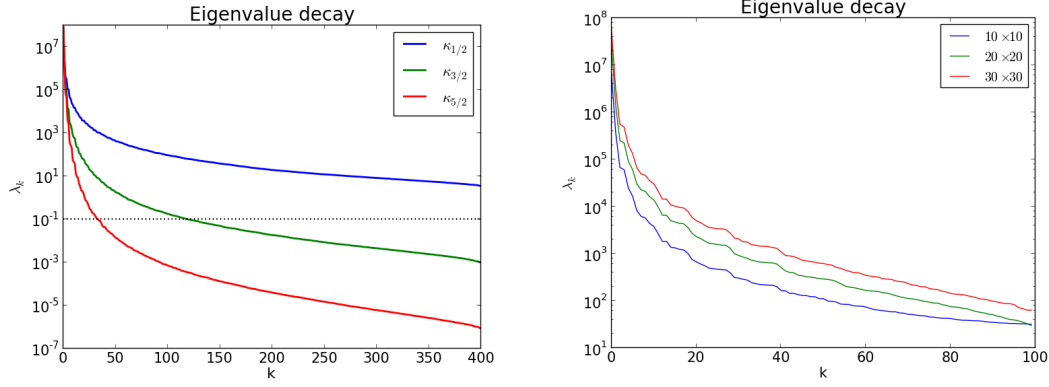


Figure 5: (left) The decay of the eigenvalues for different covariance kernels. The unknowns are discretized on a 200×200 grid and the number of measurements are $n_m = 400$. Three different covariance kernels are considered described in Equation (17) with $\theta = 10^{-3}$. (right) The decay of the eigenvalues for different setups with differing number of measurements ranging from 10×10 to 30×30 . The covariance kernel was chosen to be $\kappa_{1/2}$ described in Equation (17) and the unknowns were discretized on a 200×200 grid. Only the top 100 eigenvalues are displayed.

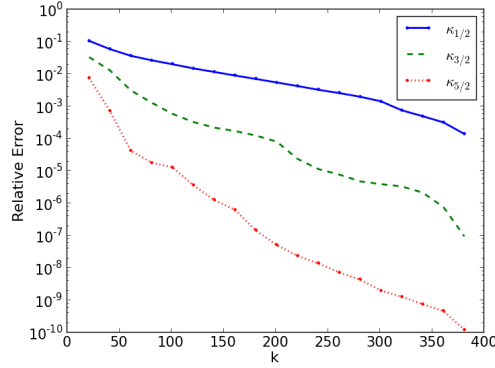


Figure 6: Dependence of the relative error in the approximation to the variance computed as a function of number of retained eigenvalues retained in the approximation. The relative error computed with the exact variance, where relative error is $\sum_k |\text{Var}_k^{\text{exact}} - \text{Var}_k^{\text{approx}}| / \sum_k |\text{Var}_k^{\text{exact}}|$

4.2.2 Cost of computing variance

Two factors are necessary to ensure that the cost of computing the variance scales linearly with the dimensionality of the unknowns parameters - the number of eigenvalues retained is independent of the mesh size and the various costs required to capture the dominant eigenmodes scales linearly with the number of unknowns.

We address the first issue, that is, the number of eigenvalues retained is independent of the mesh size. Figure 7 shows the dependence of the spectrum of the generalized eigenvalue problem $H_{\text{red}}x = \lambda \Gamma_{\text{prior}}^{-1}x$ with the dimension of the input parameter space. We use the same setup just described above, with 400 measurements. The unknown parameters, in this case the slowness of the medium, is discretized on a number of grids ranging from 100×100 to 1000×1000 . The prior is chosen as a member of the Matérn covariance class with $\nu = 3/2$ i.e. $\kappa_{3/2}$, see Equation (17). As can be seen from the figure, the number of unknowns does not affect the spectrum of the generalized eigenvalue problem and thus, does not affect the number of retained eigenvalues required to approximate the posterior covariance matrix. We address the second issue - computational costs of computing

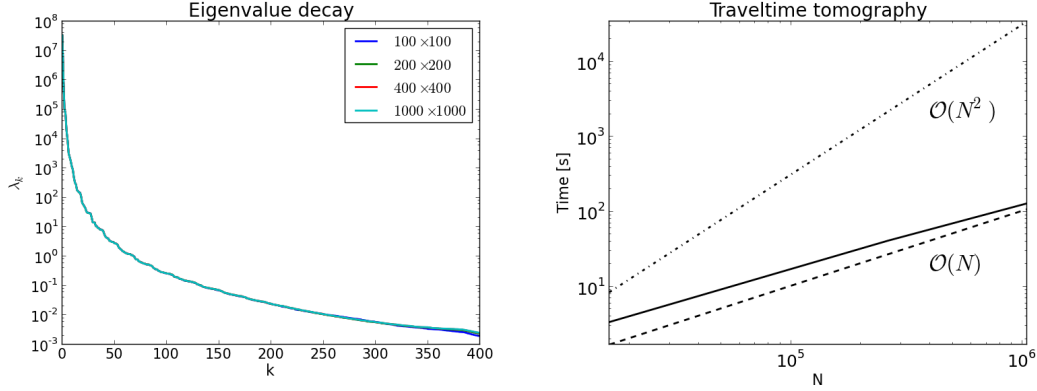


Figure 7: (left) The decay of the eigenvalues for the covariance kernel $\kappa(r) = \theta(1 + \sqrt{3}r/L)\exp(-\sqrt{3}r/L)$ where the unknowns are discretized on a number of grids ranging from 100×100 to 1000×1000 . (right) The CPU time (solid line) for computing the top 300 eigenmodes of the ray tomography example with unknowns discretized on grids ranging from 128×128 to 1024×1024 with $n_m = 1000$ measurements. The slopes of the dashed and dotted lines correspond to computational costs of $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ respectively. As can be seen, the computational cost scales linearly.

eigendecomposition. The major component in this calculation is the matvecs involving H_{red} , Γ_{prior} and $\Gamma_{\text{prior}}^{-1}$. In Section 3, we have already argued that the costs involving the aforementioned matvecs scale linearly or almost linearly with the number of unknowns. Therefore, we expect the cost of the eigendecomposition, and as a result the variance calculations, to scale similarly with the number of unknowns. This is illustrated in Figure 7. As can be seen from the figure, the computational costs scale linearly with the number of unknowns.

5 Application: Steady-state Hydraulic Tomography

Hydraulic Tomography (HT) is a technique for estimating the subsurface parameters such as transmissivity by a series of pumping tests, in which pressure (head) is measured and this data is used to reconstruct the parameters of interest by using suitable inversion algorithms. The governing equations are given by the ground water flow equations, which in 2D take the following form

$$\begin{aligned} -\nabla(K(x)\nabla\phi) &= Q(x - x_s) & x \in \Omega \\ \phi &= 0 & x \in \partial\Omega \end{aligned} \quad (18)$$

where x_s are the pumping locations, and homogenous Dirichlet boundary conditions are applied everywhere on the boundary $\partial\Omega$.

The inverse problem is to reconstruct the transmissivity field $K(x)$ from discrete measurements of ϕ obtained by repeated pumping tests at several pumping locations. In order for the system of Equations (18) to be well posed, the reconstructions of $K(x)$ need to be positive and therefore, we make a log-transformation $s(x) \stackrel{\text{def}}{=} \log K(x)$ to ensure positivity. This transformation makes the problem of reconstruction a quasi-linear inverse problem. The setup of the pumping tests is provided in Figure 8; the black squares indicate the locations at which water is pumped and the head response is computed at the receiver locations (indicated by red asterisks).

The reconstruction of transmissivity field $K(x)$ from discrete measurements of ϕ is a nonlinear ill-posed inverse problem. For nonlinear inverse problems, unlike linear inverse problems, the posterior distribution is no longer Gaussian. The measurement operator is linearized about the current

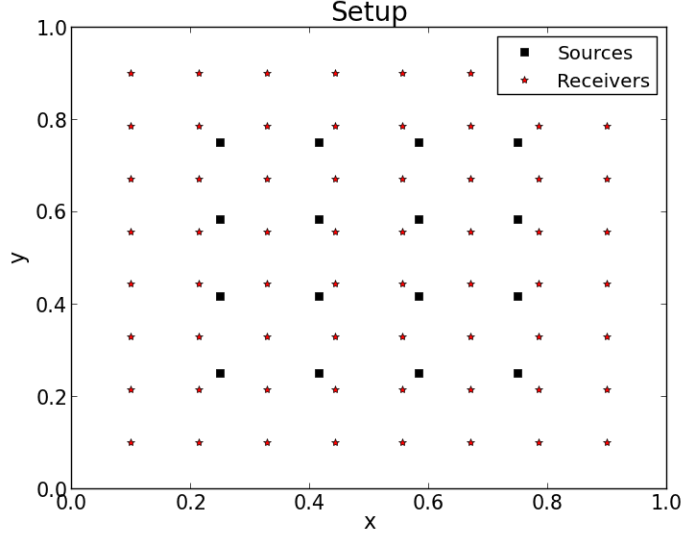


Figure 8: Setup of the 16 pumping sources (black solids) and 64 measurement locations (red asterisks). In all, there were $16 \times 64 = 1024$ measurements.

estimate s_k to obtain

$$h(s) \approx h(s_k) + J(s - s_k) \quad J = \left. \frac{\partial h(s)}{\partial s} \right|_{s=s_k}$$

The MAP point is the minimizer of the negative log-likelihood posterior PDF described in Equation (5) and can be computed using the quasi-linear geostatistical approach [25]. Essentially this involves solving a sequence of regularized linear least squares problems until the algorithm converges. The system of Equations (19) is solved using restarted GMRES. The prior covariance matrix Γ_{prior} and the Jacobian J_k are not constructed explicitly. Matvecs involving Γ_{prior} are handled in $\mathcal{O}(m \log m)$ using the FFT based approach or using \mathcal{H} -matrices, whereas the matvecs involving the Jacobian are handled using the approach described in [19]. In order to accelerate the convergence of GMRES, we use a preconditioner developed in [36]. The algorithm for computing the MAP estimate is summarized in Algorithm 2.

Algorithm 2 Quasi-linear Geostatistical Approach

- 1: **while** Not converged **do**
- 2: Solve the system of equations,

$$\begin{pmatrix} J_k \Gamma_{\text{prior}} J_k^T + \Gamma_{\text{noise}} & J_k X \\ (J_k X)^T & 0 \end{pmatrix} \begin{pmatrix} \xi_{k+1} \\ \beta_{k+1} \end{pmatrix} = \begin{pmatrix} y - h(s_k) + J_k s_k \\ 0 \end{pmatrix} \quad (19)$$

where the Jacobian is defined as $J_k = \left. \frac{\partial h(s)}{\partial s} \right|_{s=s_k}$.

- 3: The update s_{k+1} is computed by,

$$s_{k+1} = X \beta_{k+1} + \Gamma_{\text{prior}} J_k^T \xi_{k+1} \quad (20)$$

 {If necessary, add a line search}

- 4: **end while**
-

Once we have the MAP estimate, we approximate the measurement operator by a linearizing about the MAP point so as to approximate the posterior distribution about the MAP estimate by a Gaussian distribution. This is a reasonable approximation when the measurement operator is nearly

linear. This is a useful approximation even when the posterior distribution is highly nonlinear, since the Gaussian approximation is frequently used as a proposal distribution for MCMC algorithms used to explore the posterior [33]. The approximate posterior distribution $\tilde{p}(s, \beta|y)$ can be expressed as

$$\tilde{p}(s, \beta|y) \sim \mathcal{N}\left(\begin{bmatrix} \hat{s} \\ \hat{\beta} \end{bmatrix}, \Gamma_{\text{post}}\right) \quad \Gamma_{\text{post}} = \begin{pmatrix} \Gamma_{\text{prior}}^{-1} + J^T \Gamma_{\text{noise}}^{-1} J & \Gamma_{\text{prior}}^{-1} X \\ (\Gamma_{\text{prior}}^{-1} X)^T & X^T \Gamma_{\text{prior}}^{-1} X \end{pmatrix}^{-1} \quad (21)$$

where \hat{s} and $\hat{\beta}$ are the parameters evaluated at the MAP point, and J is the Jacobian calculated at the MAP point \hat{s} .

Computing the posterior covariance matrix Γ_{post} or its inverse explicitly is computationally infeasible for large scale problems. This is because (1) constructing the Jacobian explicitly requires the solution of several (possibly time-dependent) forward or adjoint partial differential equations (2) the covariance matrix is typically dense, and therefore computing matrix products and inverses involving the covariance matrix is nearly impossible in terms of storage and computational cost and finally, (3) the resulting matrix is dense and its storage can be infeasible for large problems. The approximation of the posterior covariance matrix is approximated using the same randomized approach described in Section 3.

We discuss the implementation of these ideas to a concrete application of steady-state hydraulic tomography. For the reconstruction, we use a covariance kernel, that is part of the Matérn family, and is given by

$$\kappa(x, y) = \exp\left(-\frac{\|x - y\|_2}{4L}\right) \quad (22)$$

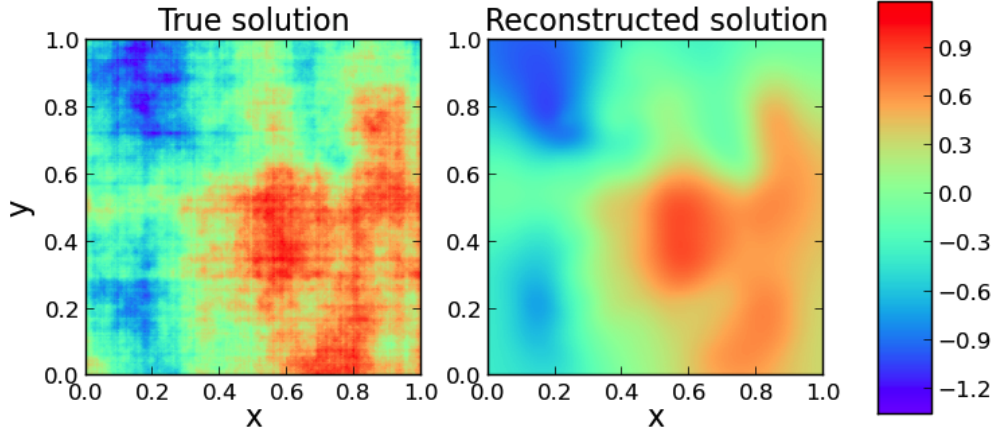


Figure 9: (left) True field which is a realization of a Gaussian process with zero mean and covariance kernel (22), and (right) reconstruction using the geostatistical approach on a grid with 401×401 points. The reconstruction error in relative L^2 sense was 0.29.

To simulate experimental error, we add Gaussian noise of 0.1% to the measurements. The MAP estimate is computed by solving the optimization problem (5) using the Gauss-Newton algorithm that was described in Algorithm 2. At each iteration the system of equations is solved using restarted GMRES (50) using the same preconditioner described in [36]. The low-rank decomposition is computed in the same fashion described in Section 4 with 30 terms in the low-rank approximation. The low-rank decomposition is computed only once; however, the preconditioner is re-built every Gauss-Newton iteration. The iterative solver took about 16 iterations on average to converge to a relative tolerance of 10^{-7} . The Gauss-Newton solver took about 6 iterations for the relative L^2 difference between subsequent iterates to go below 10^{-3} . The error in the reconstruction was 29% and the reconstruction is plotted in Figure 9.

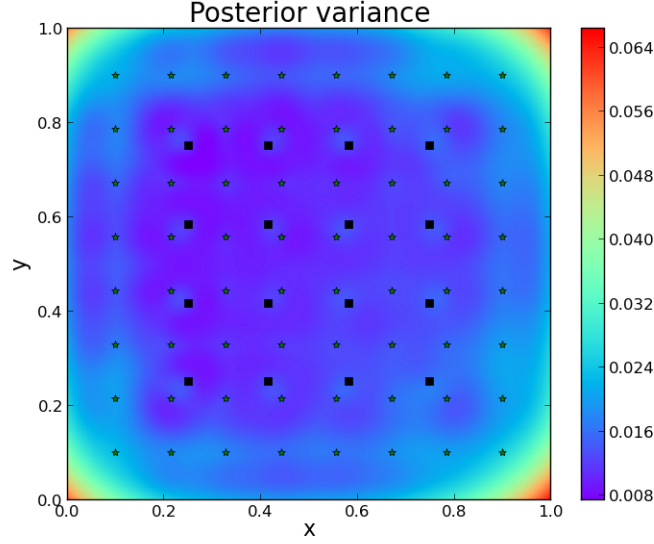


Figure 10: Visualization of the diagonals of the posterior covariance matrix in Equation (21). The computation was calculated with the Jacobian evaluated at the MAP estimate.

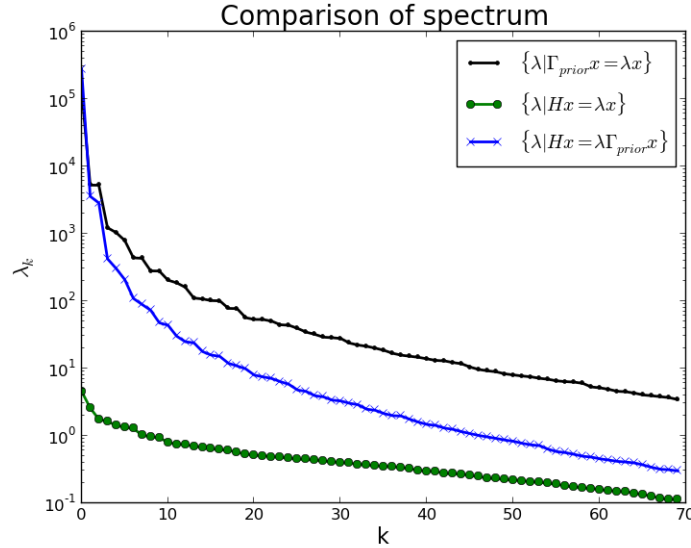


Figure 11: Comparison of the eigenvalues of the data misfit Hessian $H_{\text{red}} = J^T \Gamma_{\text{noise}}^{-1} J$, the prior covariance matrix Γ_{prior} and the generalized HEP in the Figure 9. The Jacobian is computed by linearizing the measurement operator at the MAP estimate. As can be seen, the eigenvalues of the GHEP $H_{\text{red}} x = \lambda \Gamma_{\text{prior}}^{-1} x$ decays rapidly and we can retain only the largest eigenmodes above a user-defined tolerance.

The generalized eigenvalue problem in Equation (9) is solved with the Jacobian computed at the MAP estimate and the eigenpairs corresponding to the 50 largest eigenvalues are computed using the randomized GHEP algorithm described in Section 3.2. The posterior covariance matrix is computed using Equation (21). The variance computed as the diagonals of the posterior covariance is visualized in Figure 10. The generalized eigenvalues are plotted in Figure 11. Since the true variance

is hard to compute, as it requires sampling from the true posterior distribution, comparison of the accuracy against the true variance is hard to establish. However, from the analysis in Section 4.2 the computation of F_{ss}^{-1} using the formula in (12) a cutoff of $\lambda > 0.1$ is a good approximation to the true F_{ss}^{-1} computed at the MAP estimate. As can be seen in the Figure 11, the eigenvalues of the GHEP defined in Equation (9) and the number of eigenvalues retained 70 that satisfies the cutoff $\lambda > 0.1$, is far smaller than the number of unknowns 160,801 or the number of measurements 1024. Finally, in Figure 12, we report the cost of computing the dominant eigenmodes of the GHEP defined in Equation (9). The major component in this calculation is the matvecs involving H_{red} , Γ_{prior} and $\Gamma_{\text{prior}}^{-1}$. In Section 3, we have already argued that the costs involving the aforementioned matvecs scale linearly or almost linearly with the number of unknowns. Therefore, we expect the cost of the eigendecomposition, and as a result the variance calculations, to scale similarly with the number of unknowns. This is illustrated in Figure 12. As can be seen from the figure, the computational costs scale linearly with the number of unknowns.

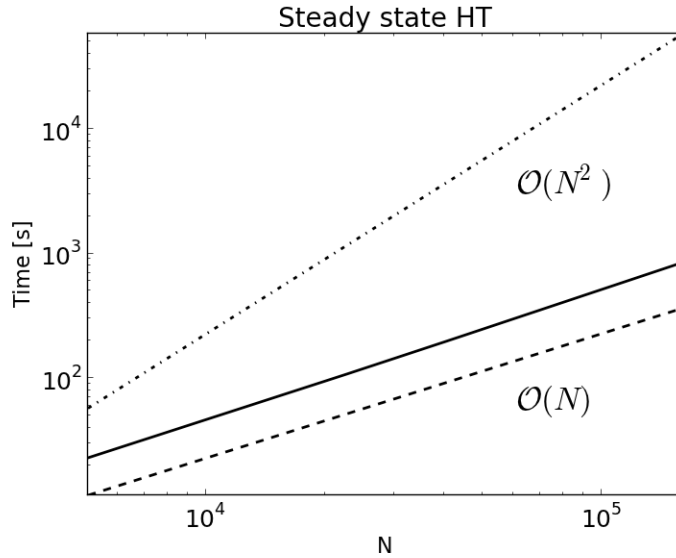


Figure 12: The CPU time (solid line) for computing the top 70 eigenmodes of the hydraulic tomography example with unknowns discretized on grids ranging from 51×51 to 401×401 grid with $n_m = 1024$ measurements. The slopes of the dashed and dotted lines correspond to computational costs of $\mathcal{O}(N)$ and $\mathcal{O}(N^2)$ respectively. As can be seen, the computational cost scales linearly. It can be seen that the computational cost scales linearly with the number of unknowns.

6 Measures of uncertainty

An experimentalist interested in designing experiments to estimate parameters is interested in obtaining as much information as possible from laborious or expensive experiments. In the context of Bayesian experimental design, one tries to optimize certain criteria based on scalar measures of conditional uncertainty. Several studies have focused their attention on optimal experimental design targeted to regression-like problems. However, the large computational costs associated with storing and computing the covariance matrices prohibit their utility in optimal design for geostatistical approach. Moreover, in contrast to regression based methods the inverse problems arising from environmental sciences are highly under-determined, i.e. $n \ll m$. As a result, several uncertainty measures defined in the geostatistical context have different theoretical properties and significance compared to the (typically) over-determined regression-based problems. For a good review of uncertainty measures in the context of the geostatistical approach, please refer to [32]. Several measures

of uncertainty are proposed based on alphabetic optimality that is prevalent in literature on regression. Because of the high computational complexity associated with storing and computing with the posterior covariance matrices, several of the optimality criteria are either computationally intensive or completely infeasible (for example, D-optimality criterion). Several of these optimality measures can be easily computed when the number of measurements are small, even when the unknowns are discretized on a very fine grid. However, as the number of measurements grow, the computational cost associated with estimating the optimality criteria can grow as either $\mathcal{O}(nm + n^3)$ in certain cases.

The optimality criteria are then based upon the invariants of the posterior covariance matrix, such as the trace, log determinant etc. When the posterior PDF is Gaussian, the posterior covariance matrix is provided by Equation (8). When the posterior distribution is strongly non-Gaussian, the approximation using (21) is no longer a good approximation to the posterior covariance and one has to resort to Monte-Carlo methods for computing these optimality criteria [24]. Our approach is to approximate the posterior covariance by computing the dominant eigenmodes of the GHEPin Equation (9) and using the approximation described in (14) and the procedure described in Section 3. Here, we describe the calculation of a few measures of uncertainty based on the posterior covariance matrix and efficient ways to calculate it. These notations are also described in [32].

1. A-optimality

$$\phi_A \stackrel{\text{def}}{=} \frac{1}{m + n_p} \text{Trace}(A\Gamma_{\text{post}}) \quad (23)$$

For $A = I$ the A-optimality criterion takes the special form $\phi_I = \frac{1}{m+n_p} \text{Trace}(\Gamma_{\text{post}})$. To compute ϕ_I , we start with the observation that

$$\text{Trace}(\Gamma_{\text{post}}) = \text{Trace}(F_{ss}^{-1}) + \text{Trace}(GS^{-1}G^T) + \text{Trace}(S^{-1})$$

Since S is of size $n_p \times n_p$ we can compute S^{-1} and its trace explicitly in $\mathcal{O}(n_p^3)$. Since G is $m \times n_p$, $\text{Trace}(GS^{-1}G^T)$ can be computed in $\mathcal{O}(mn_p)$. Finally, $\text{Trace}(F_{ss}^{-1})$ can be computed as

$$\text{Trace}(F_{ss}^{-1}) \approx \text{Trace}(\Gamma_{\text{prior}}) - \text{Trace}(U_k D_k U_k^T) \quad (24)$$

which can be computed in $\mathcal{O}(mk)$ and $\text{Trace}(\Gamma_{\text{prior}})$ is known from the covariance kernel. In summary, the calculation of $\text{Trace}(\Gamma_{\text{post}})$ is $\mathcal{O}(k + n_p)m$. In practice, n_p is $\mathcal{O}(1)$, so the terms involving it have negligible contribution to the overall computational cost.

For $A \neq I$, ϕ_A can be approximated using the Hutchinson trace estimator. Essentially,

$$\text{Trace}(A\Gamma_{\text{post}}) \approx \frac{1}{s} \sum_{i=1}^s v_i^T A\Gamma_{\text{post}} v_i$$

where the entries of v_i are chosen as ± 1 with equal probability. This forms an unbiased estimator for the trace of $A\Gamma_{\text{post}}$. However, the variance of the estimator can be large. To remedy this, other sophisticated sampling schemes for estimating the trace of a matrix, based only on the availability of fast matvecs, have been proposed in [5].

2. C-optimality

$$\phi_C \stackrel{\text{def}}{=} \frac{1}{m + n_p} c^T \Gamma_{\text{post}} c \quad (25)$$

where $c = [\partial z / \partial s^T \partial z / \partial \beta^T]^T$ is the sensitivity of a scalar model prediction z with respect to the estimated parameters, yielding the prediction variance for z . This criterion can be easily extended to vector valued model predictions as well. For details, see [32]. ϕ_C is a special case of ϕ_A . In particular,

$$\phi_C = \frac{1}{m + n_p} \text{Trace}(cc^T \Gamma_{\text{post}}) = \phi_{A=cc^T}$$

Given the vector c , it is easy to see that ϕ_C can be calculated in $\mathcal{O}(m \log m + km)$. Furthermore, c can be computed efficiently using adjoint-based techniques [32].

3. D-optimality

$$\phi_D \stackrel{\text{def}}{=} \log \det(\Gamma_{\text{post}})$$

For numerical reasons, the logarithm is considered here instead of the determinant itself. The calculation of the determinant is computationally infeasible. However, since Γ_{post} is a symmetric positive definite matrix

$$\log \det(\Gamma_{\text{post}}) = \text{Trace}(\log \Gamma_{\text{post}})$$

Then, in a procedure similar to the A -optimality criterion, we can compute the trace of the matrix using the Hutchinson trace estimator

$$\text{Trace}(\log \Gamma_{\text{post}}) \approx \frac{1}{s} \sum_{i=1}^s v_i^T \log \Gamma_{\text{post}} v_i$$

As before, the entries of v_i are chosen as ± 1 with equal probability. Calculating the entries of $\log \Gamma_{\text{post}}$ is also computationally infeasible. However, matvecs of the form $\log(\Gamma_{\text{post}})v_i$ can be formed efficiently using matrix-free techniques described in [14, 20]. When the posterior PDF is Gaussian, i.e. for linear inverse problems, the D-optimality criterion is closely related to maximizing the entropy of the random variable. The entropy of a random variable X with PDF $p(X)$ is defined as $H[X] = E[-\log p(X)]$, where $E[\cdot]$ is the expectation. For Gaussian distributions $X \sim \mathcal{N}(\mu, \Sigma)$, the entropy can be calculated as $H[X] = \frac{1}{2} \log 2\pi e + \frac{1}{2} \log \det(\Sigma)$. Therefore, $\phi_D = 2H[s, \beta|y] - \log 2\pi e$.

An alternative way to compute D-optimality is to use the properties of determinants - the determinant of the product of matrices is the product of the determinants. Applying this result to Equation (14), we have that

$$\begin{aligned} \log \det(\Gamma_{\text{post}}) &= \log \det(F_{ss}^{-1}) + \log \det(S^{-1}) \\ &\approx \log \det(\Gamma_{\text{prior}}) + \log \det(I - D_k) + \log \det(S^{-1}) \end{aligned} \tag{26}$$

This result follows because

$$\begin{aligned} \log \det(F_{ss}^{-1}) &\approx \log \det(\Gamma_{\text{prior}}) + \log \det\left(I - \Gamma_{\text{prior}}^{-1} U_k D_k U_k^T\right) \\ &= \log \det(\Gamma_{\text{prior}}) + \log \det\left(I - U_k^T \Gamma_{\text{prior}}^{-1} U_k D_k\right) \\ &= \log \det(\Gamma_{\text{prior}}) + \log \det(I - D_k) \end{aligned}$$

where the equality in the second line followed from a straightforward application of the Sylvester's determinant lemma followed by the use of the fact that the generalized eigenvectors U_k were $\Gamma_{\text{prior}}^{-1}$ -orthonormal. Further, we can treat $\log \det(\Gamma_{\text{prior}})$ as constant because it does not change with the measurements.

4. E-optimality

$$\phi_E \stackrel{\text{def}}{=} \max \text{eig}(\Gamma_{\text{post}})$$

This is related to the C-optimality as $\max_{\|c\|_2=1} \phi_C$. The resulting vector c that maximizes the C-optimality criterion is the direction of the largest variation. Computing the largest eigenvalue (and eigenvector) can be easily computed by using a few matvecs involving Γ_{post} using few iterations of either the power iteration or Lanczos method. It can also be efficiently estimated using the randomized GHEP algorithm that we describe in Section 3.2 with $A = \Gamma_{\text{post}}$ and $B = I$.

To illustrate these ideas, we consider again the ray tomography example in Section 4. We choose two different experimental setups to compare the uncertainty measures described in Section 6. Both setups have the same number of sources 20 and same number of receivers 20 discretized on a domain of size 512×512 . However, they differ in the locations of the sources and as a result produce different reconstructions of the same field, see Figure 13. But a natural question to ask is: which experimental setup is better? Intuitively, one would expect that when the sources cover a larger area the reconstruction would be better over a larger area and hence, the resulting uncertainty (measured as a function of the posterior covariance) would be lower in the entire domain. Indeed, this is the case. Quantitatively, the variance of setup 2 is larger than that of setup 1 and this is reflected in all the measures of uncertainty. This is listed in Table 1.

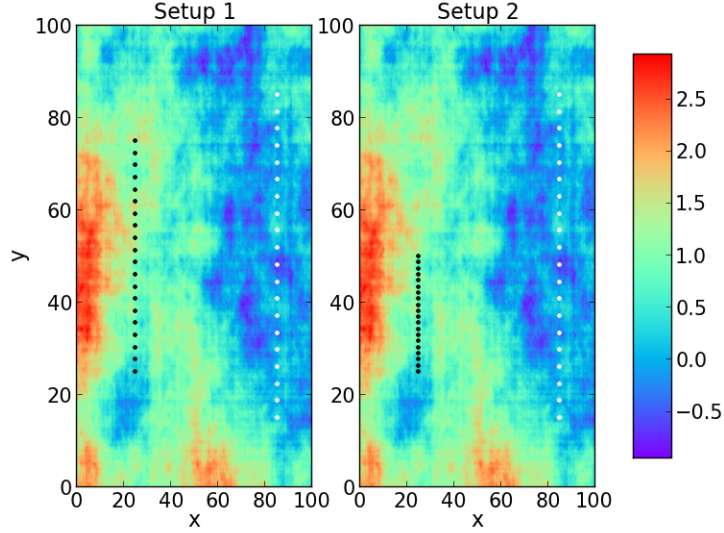


Figure 13: Two different setups for comparing uncertainty measures defined in 6. Black dots correspond to sources and white dots correspond to receivers, of which there are 20 each. The true field is discretized on a 512×512 grid and is a realization of the covariance kernel given in (16).

Setup	ϕ_A	ϕ_C	$\tilde{\phi}_D$	ϕ_E	$\text{Trace}S^{-1}$
1	64.64	5.22	-823.18	26.28	6.31×10^{-4}
2	72.33	6.27	-688.81	26.50	6.46×10^{-4}

Table 1: Comparison of the various measures of uncertainty defined in Section 6 applied to the setups described in Figure 13.

For the computation of ϕ_A , we chose $A = I$, whereas for computing ϕ_C we chose c as a vector of ones. We observe that $\tilde{\phi}_D$ is the most sensitive measure of uncertainty because the difference in $\tilde{\phi}_D$ between the two setups is the largest, even in relative magnitude. The least sensitive is the E-optimality criterion which only takes into account the largest eigenmode. ϕ_A measures the trace of the posterior covariance matrix. Both ϕ_A and ϕ_D account for both large-scale and small-scale variations since it takes into account all the eigenmodes of the posterior GHEP in Equation (9). While computing $\tilde{\phi}_D$ it is required to compute $\log \det(I - D_k)$, and ϕ_A requires weighted sum of the D_k). Therefore, small changes made to the lowest eigenmodes (corresponding to fine-scale variations) have a higher impact in the calculations of ϕ_D , rather than ϕ_A . By the special choice of the vector c , which is chosen to be a vector of ones, ϕ_C is proportional to the variance of predicting the global mean and can also be interpreted as proportional to the average conditional integral scale [32]. Since in setup 1 the sources are spread over a larger area and captures the large scale variability well but

has difficulty with small scale features, leading to smaller integral scale and lower ϕ_C . By a similar reasoning $\text{Trace}(S^{-1})$, which represents the uncertainty in the parameters β , is larger for setup 2.

7 Conclusions and Future Work

Several papers have focused their attention on solving inverse problems. The problem of quantifying the uncertainty associated with the estimate is challenging because of the high dimensionality of the space of unknowns and the high computational costs involving the representing unknown random fields and computing the solution to the forward . We consider the problem of quantifying the associated uncertainty by computing an approximation to the posterior covariance matrix which is roughly composed of two terms - the prior covariance matrix and a low rank term that contains the dominant eigenmodes of a generalized Hermitian eigenvalue problem (GHEP) involving the misfit portion of the Hessian and the inverse prior covariance matrix. For several inverse problems, the spectrum of the eigenproblem, described above, decays rapidly and we are justified in retaining only the largest eigenvalues satisfying a cutoff $\lambda > \varepsilon$. We have provided an efficient algorithm for computing this low-rank representation using a randomized algorithm that deliberately avoids expensive computations involving square-root (or its inverse) of the prior covariance matrix. The accuracy of the low-rank representation and the parameters controlling the number of retained eigenvalues have been explored. Finally, we have shown how to approximately compute measures of uncertainty based on the approximate representation of the posterior covariance matrix. The resulting algorithms are highly scalable and their performance has been demonstrated on two challenging applications - ray based tomography and steady-state hydraulic tomography.

For nonlinear problems we have used the approximate posterior distribution which is approximated to be a Gaussian by linearizing the measurement operator at the MAP point. However, for highly nonlinear problems this approximation may not be reasonable. In such cases, one may have to resort to sampling from the full posterior in order to quantify the uncertainty [29]. However, such a process would be challenging for much the same reasons - high dimensionality of the input spaces and the high computational costs associated with the forward problem. A limitation of our analysis is that it is restricted to the case for which the priors can be represented as a Gaussian random field. Future work is necessary to extend our analysis to different priors.

In future, we would like to explore the possibility of using the uncertainty measures described in Section 6 to optimize the control variables in an experimental setup in order to obtain the most amount of information from laborious or expensive experiments. Finally, another challenging application that would benefit from an efficient representation of posterior covariance, is data assimilation using Kalman filter. Implementing Kalman filter can be computationally expensive because of the high computational costs in updating the posterior distribution. For a certain class of problems, we anticipate that the posterior covariance matrix can be represented as the combination of the prior and a low-rank term, that can be recursively updated efficiently. This could lead to a highly scalable implementation of the Kalman filter. This will be discussed in subsequent papers.

8 Acknowledgements

The major work on this paper was completed when the first author was a PhD candidate in ICME Stanford University and he would like to thank Ivy Huang for all her help and support during this process and beyond. The authors were supported by NSF Award 0934596, Subsurface Imaging and Uncertainty Quantification. We thank Tania Bakhos for a careful reading of the manuscript and providing valuable suggestions to improve readability.

References

- [1] Alen Alexanderian, Philip Gloor, and Omar Ghattas. On bayesian a-and d-optimal experimental designs in infinite dimensions. *arXiv preprint arXiv:1408.6323*, 2014.

- [2] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A-optimal design of experiments for infinite-dimensional bayesian linear inverse problems with regularized ℓ_0 -sparsification. *SIAM Journal on Scientific Computing*, 36(5):A2122–A2148, 2014.
- [3] Alen Alexanderian, Noemi Petra, Georg Stadler, and Omar Ghattas. A fast and scalable method for a-optimal design of experiments for infinite-dimensional bayesian nonlinear inverse problems. *arXiv preprint arXiv:1410.5899*, 2014.
- [4] Sivaram Ambikasaran, Judith Yue Li, Peter K Kitanidis, and Eric Darve. Large-scale stochastic linear inversion using hierarchical matrices. *Computational Geosciences*, 17(6):913–927, 2013.
- [5] Haim Avron and Sivan Toledo. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM (JACM)*, 58(2):8, 2011.
- [6] Zhaojun Bai, James Demmel, Jack Dongarra, Axel Ruhe, and Henk Van Der Vorst. *Templates for the solution of algebraic eigenvalue problems: a practical guide*, volume 11. Society for Industrial and Applied Mathematics, 1987.
- [7] R.K. Beatson, J.B. Cherrie, and C.T. Mouat. Fast fitting of radial basis functions: Methods based on preconditioned GMRES iteration. *Advances in Computational Mathematics*, 11(2):253–270, 1999.
- [8] M. Bebendorf. Hierarchical LU decomposition-based preconditioners for BEM. *Computing*, 74(3):225–247, 2005.
- [9] T. Bui-Thanh and O. Ghattas. Analysis of the Hessian for inverse scattering problems: I. Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):055001, 2012.
- [10] T. Bui-Thanh and O. Ghattas. Analysis of the Hessian for inverse scattering problems: II. Inverse medium scattering of acoustic waves. *Inverse Problems*, 28(5):055002, 2012.
- [11] Tan Bui-Thanh, Carsten Burstedde, Omar Ghattas, James Martin, Georg Stadler, and Lucas C Wilcox. Extreme-scale UQ for Bayesian inverse problems governed by PDEs. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, page 3. IEEE Computer Society Press, 2012.
- [12] Tan Bui-Thanh, Omar Ghattas, James Martin, and Georg Stadler. A computational framework for infinite-dimensional Bayesian inverse problems part i: The linearized case, with application to global seismic inversion. *SIAM Journal on Scientific Computing*, 35(6):A2494–A2523, 2013.
- [13] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.
- [14] Jie Chen, Mihai Anitescu, and Yousef Saad. Computing $f(A)b$ via least squares polynomial approximations. *SIAM Journal on Scientific Computing*, 33(1):195–222, 2011.
- [15] G. Christakos. On the problem of permissible covariance and variogram models. *Water Resources Research*, 20(2):251–265, 1984.
- [16] CR Dietrich and Garry Neil Newsam. Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107, 1997.
- [17] CR Dietrich and GN Newsam. Efficient generation of conditional simulations by chebyshev matrix polynomial approximations to the symmetric square root of the covariance matrix. *Mathematical geology*, 27(2):207–228, 1995.
- [18] H.P. Flath, L.C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas. Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432, 2011.

- [19] E. Haber, U.M. Ascher, and D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse problems*, 16(5):1263, 2000.
- [20] Nicholas Hale, Nicholas J Higham, and Lloyd N Trefethen. Computing A^α , $\log(A)$, and related matrix functions by contour integrals. *SIAM Journal on Numerical Analysis*, 46(5):2505–2523, 2008.
- [21] N. Halko, P.G. Martinsson, and J.A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.
- [22] V. Hernandez, J. E. Roman, A. Tomas, and V. Vidal. A survey of software for sparse eigenvalue problems. Technical Report STR-6, Universitat Politècnica de València, 2009. Available at <http://www.grycap.upv.es/slepc>.
- [23] Dionissios T Hristopulos. Spartan Gibbs random field models for geostatistical applications. *SIAM Journal on Scientific Computing*, 24(6):2125–2162, 2003.
- [24] Xun Huan and Youssef M Marzouk. Simulation-based optimal Bayesian experimental design for nonlinear systems. *Journal of Computational Physics*, 2012.
- [25] P. K. Kitanidis. Quasilinear geostatistical theory for inversing. *Water Resour. Res.*, 31(10):2411–2419, 1995.
- [26] P. K. Kitanidis. *On stochastic inverse modeling*, volume Geophysical Monograph 171, pages 19–30. AGU, Washington, D. C., 2007.
- [27] P. K. Kitanidis. *Bayesian and Geostatistical Approaches to Inverse Problems*, pages 71–85. John Wiley & Sons, Ltd, 2010.
- [28] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.
- [29] J. Martin, L.C. Wilcox, C. Burstedde, and O. Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):1460–1487, 2012.
- [30] G. Matheron. The intrinsic random functions and their applications. *Advances in applied probability*, 5(3):439–468, 1973.
- [31] W. Nowak, S. Tenkleve, and O.A. Cirpka. Efficient computation of linearized cross-covariance and auto-covariance matrices of interdependent quantities. *Mathematical Geology*, 35(1):53–66, 2003.
- [32] Wolfgang Nowak. Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design. *Mathematical Geosciences*, 42(2):199–221, 2010.
- [33] Noemi Petra, James Martin, Georg Stadler, and Omar Ghattas. A computational framework for infinite-dimensional Bayesian inverse problems: Part ii. Stochastic Newton MCMC with application to ice sheet flow inverse problems. *arXiv preprint arXiv:1308.6221*, 2013.
- [34] Youcef Saad. *Numerical methods for large eigenvalue problems*, volume 158. SIAM, 1992.
- [35] AK Saibaba, S Ambikasaran, J Yue Li, PK Kitanidis, and EF Darve. Application of Hierarchical matrices to linear inverse problems in geostatistics. *Oil and Gas Science and Technology-Revue de l’IFP-Institut Francais du Petrole*, 67(5):857, 2012.
- [36] A.K. Saibaba and P.K. Kitanidis. Efficient methods for large-scale linear inversion using a geostatistical approach. *Water Resources Research*, 48(5):W05522, 2012.

- [37] A.K. Saibaba and P.K. Kitanidis. Randomized square-root free algorithms for generalized Hermitian eigenvalue problems. *ArXiv preprint <http://arxiv.org/abs/1307.6885>*, 2014.
- [38] C. Schwab and R.A. Todor. Karhunen-Loève approximation of random fields by generalized fast multipole methods. *Journal of Computational Physics*, 217(1):100–122, 2006.
- [39] Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.
- [40] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.